A Novel Method for Detecting Similar Microblog Pages based on Longest Common Subsequence

Wengang Zhou[1,*],Huiling Guo[1],Yang Zhao[2]

1 *School of Computer Science and Technology*
ZhouKou Normal University
Zhoukou, Henan, China
2 *Department of Electronic and Information Technology*
Jiangmen Polytechnic
Jiangmen, Guangdong, China

***Abstract* — Microblog is a relation-based platform for sharing, spreading and acquiring information and it is also the source of internet public opinion and the important battlefield of information transmission. The convenient forwarding operations of microblog result in the rapid spread of plenty of identical or similar microblog pages in the microblog space. Therefore, the detection of similar microblog pages is of great importance to lighten the client's burden of browsing and improve the analytic efficiency of internet public opinion. A method based on LCS(Longest Common Subsequence) is introduced to detect similar microblog page: First is to calculate the files' subset of the possibly similar microblog pages, and next is to calculate its LCS and extract the reliable parts so as to ultimately detect the similar microblog pages. Experiments show that this method can detect the similar pages from the microblog data accurately and efficiently.**

Keyword - longest common subsequence; near-duplicate detection; similarity measurement; microblog page

I. INTRODUCTION

Microblog is a platform based on relation for sharing, communicating and access to information. Users can utilize the WEB, WAP and other client with 140 characters of text to make a comment, forward and realize the real-time share about people or events[1], so microblog with a very important position in the network of public opinion is a source initiated the dissemination of information. Microblog information has become an important data source of network public opinion browsing and analysis.

In the use of microblog, convenient forwarding operation, making a large number of the same or similar microblog page quickly spread in the microblog space. For microblog users browse, although many see the microblog information, but meaningful amount of information is very limited; for public opinion to browse and analysis, in the massive microblog information, the same or similar microblog page only has statistical meaning[2]. Therefore, the detection of similar pages on microblog is of great significance to reduce the burden of users and improve the efficiency of network public opinion analysis.

For microblog similarity detection, so far there are many researchers put forward many methods and the representative method is "based on content analysis is similar to the microblog double detection and filtering", it is believed that a large number of forwarding microblog more in similar time, the repetition rate decreases with the increase of time, use VSM model, vector angle cosine similarity value of two tweets, the value is higher than a threshold is defined as repetition, repetition rate is the total period of repeated microblog number ratio, microblog in a short time and high repetition rate, with the increase of time lag, the repetition rate decreased rapidly to microblogging a very small value. Based on the proposed content-based similar calculation of double filtering method: first to grab a time period of microblog in the first filter, subsection filtration, then the double filtering to the similar time published microblog-index filter to microblog text flow filter as a whole, such time separated by a short microblog to weight, which can not only guarantee the accuracy but also greatly reduce processing time and improve availability [3].

Considering to microblog is similar to most of the page is forwarded by a user caused, we try to use the calculation which may be similar to the twitter page document subset, and then on the basis of extraction part of the trusted, eventually detect microblog similar pages. This paper presents a LCS (Largest Common Subsequence) based on the microblog similar page detection algorithm, it can be a good measure of the similarity between the micro blog page and contains the relationship to obtain a high degree of accuracy. In this paper, a comprehensive experiment is carried out on the data from Sina microblog. The results show that the method proposed in this paper is feasible and effective.

II. RELATED RESEARCH

International for microblog similar document detection research is still relatively small, at present international is mainly for document similarity detection research, mainly for large file system. Later by the expansion application in search engine field detection of similar pages, scam systems such as Stanford University. To now, many researchers in the field put forward many methods. These methods mainly

from two aspects to distinguish: Strategy of document feature extraction from the document, and by these characteristics calculation strategy document signature. In the first aspect, the shingles and document vectors are the most commonly used features, and other Brin in the COPS system to the sentence as the unit to get shingles, Broder and other using the word particle size on the sliding window to get shingles[5]. Hoad and Chowdhury, respectively, based on document vectors developed the similarity measure and the dictionary to improve document vectors[6]. In the second aspect, based on shingles method usually use the method or methods from their characteristics to obtain the document signature. And Simhash I-match algorithm and Chowdhury, and others in the document vector algorithm to calculate the signature on the document[7]. In these techniques, shingling Charikar algorithm and Simhash Broder algorithm are considered to represent the current level of technology development, and have been used in the actual search engine system.

A. Research on the Correlation of Shingling Algorithm

Shingling Broder algorithm with two pages of resemblance and containment to measure their similarity[8]. Resemblance web page A and B is located in the interval [0,1] a numerical, is closer to 1, the two page is similar; similar, containment is a numerical value in the interval [0,1], closer to 1, explain the higher degree of a web page by page B contains.

Schleimer the winnowing algorithm on the classic shingling algorithm is improved, in order to detect two page matching part, the main process is: will the division growth into k segments, each w fragment is a window, to take the smallest fragment in each window hash values as the window of the hash value[9]. All selected hash values are used as the fingerprints of the article. The algorithm can ensure that the length of the matching string is detected with a length greater than or equal to w+k-1.

Fetterly et al. Improved shingling algorithm in the process of group evolution of similar web pages.

B. Research on the Correlation of Simhas Algorithm

Charikar Simhash (similar to hash) algorithm by comparing two pages of hash value in the same position accounted for the proportion of to measure the similarity is the main idea: page of each token mapping to a B - dimensional vector space, each dimension of the value 1 or -1. Adding all the token mapping in the page to the B dimension matrix of the page. Each non negative element in the matrix is set to 1, or 0, so that the unique hash value of the page is obtained. The hash value has the property that the similarity degree of the two pages is proportional to the number of bits in the hash value of the two pages.

Manku is designed to solve the problem of Hamming distance method, namely from a bit of the fingerprint collection quickly find with different fingerprint of a given number up to all the fingerprint, which makes similar hash in the huge web page similarity detection become practical technology.

C. Related Evaluation Work

Henzinger the shingling algorithm and Simhash algorithm effect of large-scale evaluations and approximately equal to recall rate, based on comparing the precision by adjusting the parameters. The experiment shows due to the influence of the template, the two algorithms find similar sites with similar web pages can obtain ideal effect. In the search in different sites similar pages with high precision.

D. Other Technologies in Recent Years

In the work of Yang and Callan, the detection problem of similar web pages is regarded as the restricted clustering problem at the instance level, and the semi supervised clustering algorithm is developed to solve this problem. They considered three types of constraints: must-link (recognized as belonging to the same class), cannot-link (being recognized as belonging to different classes) and family-link (possibly belonging to the same class) et al. Huffman will be similar to the detection problem as a search evaluation of a part of the study, through a variety of types of signals for each document to improve the accuracy of the signal.

III. SIMILARITY MEASURE METHOD BASED ON LCS

LCS is the abbreviation of Longest Common Subsequence, that is, the longest common sub sequence. A sequence, if two or more sequences of a sequence is known and it is the longest of all the sub sequences, the longest common sub sequence.

Solving the two sequences A and B LCS has a lot of algorithms, the typical, the best performance is proposed by Myers O (ND) algorithm. In the Myers algorithm, it solves the LCS problem that is in fact for editing map contains most diagonal path. Such as the two sequences are: A = B, abcabba = cbabac. Solving the shortest edit script (SES) from A to B is the “2D 3Ib 6D 7Ic 1D”, which can easily get the A and LCS B: caba.

Based on the Myers algorithm for solving LCS, we can propose to judge the similarity of the two documents of the metric method: if there are two documents A and B, |LCS| for their LCS length, |SES| for the shortest edit script length. Assuming A = B, cbabac = A, abcabba and B are expressed as |A| and |B|, respectively,

$$|A| + |B| = 2|LCS| + |SES| \quad (1)$$

Define rate resemble as:

$$R(A, B) = |LCS| / (|A| + |LCS| - |B|) \quad (2)$$

Define rate contain as:

$$C(A, B) = |LCS| / |B| \quad (3)$$

These two parameters respectively represent the similarity and the inclusion rate of two documents B and A. So, to judge the similarity of the two documents is: if R (A, B) or C (A, B) more than a certain threshold.

The above analysis shows that basing on LCS is similar to the general process of detection: first according to the certain method for detection of the document to sort, as far as possible to form a maximal subset of the document in the

front row; secondly read the first document, put the first subset. Finally, in order gradually to read other documents, compared with existing subset of the first document, if it was not similar generates a new subset.

In the similarity comparison of microblog pages, the order of words is an important information, LCS can well reflect the order of words. The forward operation of microblog is easy to generate a large number of pages with the relationship, through the LCS similarity detection, still think they are similar. Microblog page there are many such as advertising information of noise information. They mainly exist in the head and tail part of the page, through different locations of the words gave the corresponding weights, the LCS detection can separate this information and the page of text information, so as to improve the detection accuracy.

So LCS similarity detection algorithm mainly solves two problems based on, one is before the calculation of the LCS in as far as possible to exclude those who do not need to complete the judgment documents, only in the two documents are likely to be similar when only need again LCS calculation; second, when checking in, how to eliminate interference of microblog page, in LCS extraction truly believable to computing the document similarity.

It can be seen that the use of LCS based detection algorithm and the similarity measurement method based on this not only can get higher detection accuracy, but also can get a relatively high efficiency.

IV. DESIGN OF PREDICTIVE CONTROLLER

For effectively computing the similarity of microblog page, according to the similar detection of the measurement methods and the need to address problems based on LCS, first of all to microblog page document set partitioning for a possible similar subsets of a document, only in the document may be similar to compute the LCS; secondly, calculate the LCS and select trusted parts of it to calculate the similarity.

A. Calculation May be Similar to the Microblog Page Document Subset

According to the method mentioned above, firstly, we want to segment the collection of microblog page documents into a similar subset of documents, and then compare the similarity of the documents. Here we can use the proposed Myers O (ND) algorithm. Its steps are as follows:

(1) select the microblog pages that need to be detected, remove all the HTML tags and page formatting information, and divide the page documents into a set of sentences according to the punctuation marks.

(2) calculate the microblog page document of fingerprint, the method is: first, choose length is greater than a certain value X of the sentence, and then to calculate the MD5 digest, to K modulo, different values are divided into k groups; secondly, from each group select MD5 value distribution of the minimum in the different position of Y different words to calculate a 128 bit hash value, known as fingerprint, so there are k fingerprint; finally, the MD5 value of cyclic shift on a byte, repeat the above two steps, because the MD5 value of 16 bytes, so up to 16*k a fingerprint.

For each microblog page document, according to the steps in front of the knowledge, can get up to 16 K fingerprint. Adjust the X, Y, Z three parameters (covering Z Y a different sentence length of the text), when the parameter value is greater, indicating that the more stringent conditions, when the parameter value in a certain time, as long as the two posts page document exist with a fingerprint that can think they are similar pages of the probability is very high. So, as long as there is a fingerprint of the same, we can think that they are similar to the page. Only through a fingerprint to determine whether similar, may be missing a lot of similar micro blog page, so we can use more than one fingerprint, such as k=5, can have a maximum of 80 fingerprints. Because the microblog page is relatively simple, so the value is entirely reasonable.

According to the previous fingerprint, we can build the index of the fingerprint, and gather them together into a similar subset. Due to the microblog forwarding operation's sake, microblog page similarity exist obvious transitivity, so we can be on microblog page document sorted similarity only from large to small one-way transmission, and then put them all together to similar subset .

To sum up, the microblog page document subset of calculating the time aggregation, first the document as it can generate the number of fingerprints from big to small order, fingerprint more likely to find more similar document page, and then sequentially from twitter page document collections in one by one to read, according to the fingerprint indexing is read from the rest of the twitter page document collection and its possible similarities and according to the transfer of similar document pages together to form may similar subset.

B. Calculate LCS and Extract the Trusted Part

Microblog page noise information (for example, advertising information, etc.) usually occurs in the head and tail of the microblogging web content, two different user's Twitter page, the template is the same, and text content is generally continuous. Therefore, there are two characteristics of a microblog page features: first, the content of a trusted part of the text is continuous; two is the trusted part of the microblog page is in the central part of the.

Calculating the similarity of the microblog page document needs to calculate LCS and extract its trusted part, here, we still use the Myers O (ND) algorithm. In fact, we are here to calculate the two microblog LCS and SES documents are not strictly, but their approximate value.

Next, we focus on how to extract the trusted part of LCS, Figure 1 describes how to calculate the basic region, Figure 2 describes how to extract the trusted part from the base area.

Algorithm through the microblog page document A is divided into m=[size (A) /s] size of the s block, the n= (m2+m) /2 may be the region. Then to each block gives certain weights. Finally, in R_B can get two maximum weight and cover blocks minimum region R_{ij}.

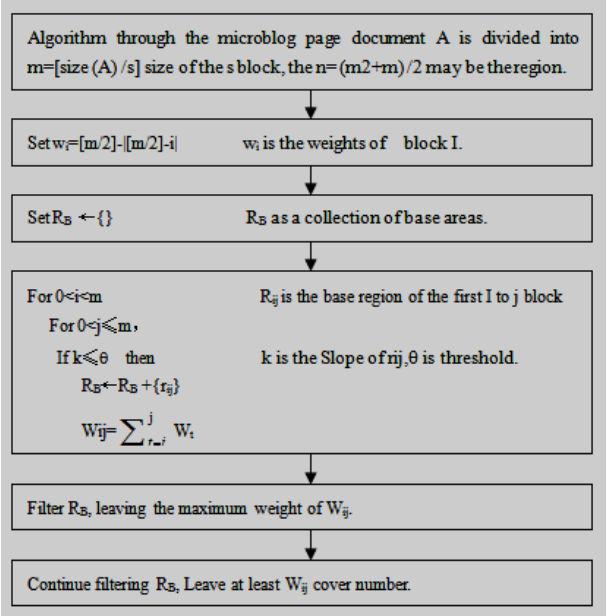

Figure 1. Calculation base area

From Figure 2 we can know, from the microblog page document based regions R_{ij}, from (i-1) * 1 to each position I * S-1 to find the first (minimum) position X_1 to make the left end point of the regional expansion to the X_1 slope area just below the threshold θ , using the same method will the right endpoints of regional expansion to the maximum position of X_R makes the regional slope just below the threshold θ. (x_r-x_1) to calculate the maximum value (i.e. the longest extension area) as the approximate credible area.

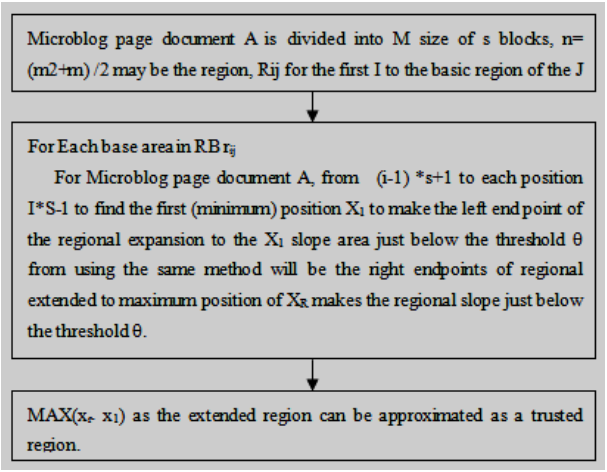

Figure 2. Calculated and extracted from the trusted zone

V. EXPERIMENT

Microblog page is created by the user, generally has a continuous text, so we will microblog page similarity relation is divided into three kinds: similar and dissimilar, pending (usually detection can not determine the). We can conduct three experiments to organize: Experiment 1, evaluate the fingerprint on the number of generated may be similar to microblog page document subset; Experiment 2 assessed credible LCS regional slope threshold effect; Experiment 3, evaluation algorithm performance.

C. Experiment 1: Assessing the Impact of the Number of Fingerprints on the Generation of a Subset of documents that May be Similar to the Microblog Page

Random from the public hall to download the 1000 microblog page document evaluation. First, for each selected twitter page document, we use the method of the collection runs a similarity detection. By calculating their LCS and extract part of the trusted, to determine whether there are similar, then we by manual inspection to judge whether similar. Secondly, for each selected micro blog page document, we can use the method to detect the possible similarity after the formation of the.

From the above algorithm, we know that the number of sets of fingerprints for different values will be different may be similar to the subset. So we have to increase the number of fingerprints from 5 to 80 for the experiment, in order to increase the value of 15, a total of 6 sets of data. Experimental results show that with the increase of the number of fingerprints, coverage rate has continued to rise, and credible rates continue to decline. That is to say credible rate is low, each microblog page need to compare the average number of times more, computational cost is higher, so you can by increasing computing resources to achieve higher coverage. Coverage and confidence are balanced when the number of fingerprints is 25.

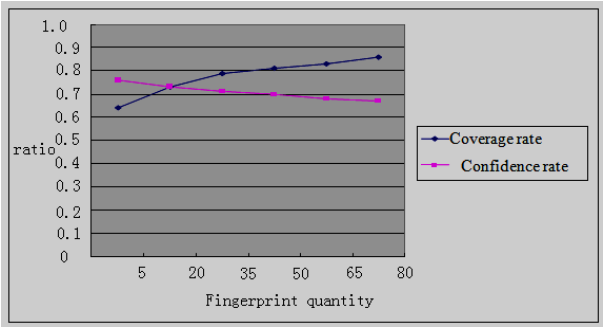

Figure 3. The effect of the number of fingerprints on the generation of the micro blog page document subset

D. Experiment 2: Evaluating the Influence of Sliding Window Length

In the generated document filtering framework, we also have two objectives: To compare the two microblog document, 1) because most of the LCS calculation (complexity is O (ND), the smaller the differences between the documents, calculation speed faster and faster, so differences in both document points in LCS calculation should be possible to be filtered out; 2) for two real similar documents (true-similar), which belongs to the TLCS part

should be do be preserved. We used rate strainaway (removal rate) and retention (rate reserved) to evaluate these two goals. Removal rate (strainaway rate, and the) definition: when two microblogging document is not really similar document not-true-similar, before they are filtered out by the text length and text content total length (filtering) ratio. Retention rate (reserved rate) is defined as: for two real similar microblogging document true-similar of document filtering framework calculated the length of the TLCS and calculated on the original document TLCS length ratio.

First, we use 112 fingerprints in the first step of 430 million document collection calculation may get a similar set, 46 million similar subset. Then, we all 430 million 100 documents from the collection of documents selected at random, and find their corresponding similar subset, as the objects of evaluation. For each selected documents, we will it possible similar sub from other documents identified by manual inspection is really similar true-similar or not really similar (not-true-similar). The two types of documents and then used to calculate the retention rate (reserved rate) and the removal rate (strainaway rate).

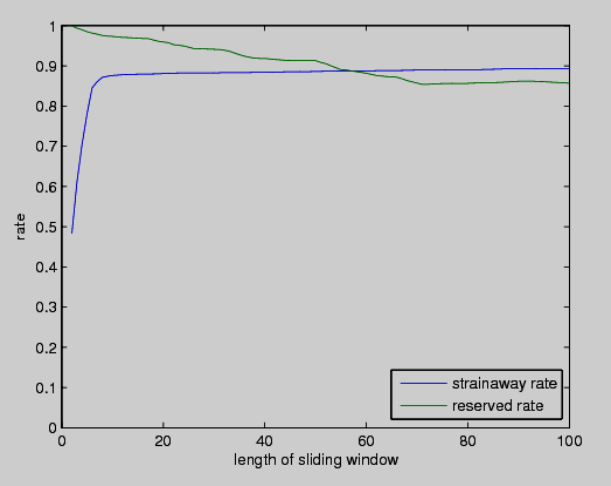

Figure 4. Evaluating the influence of sliding window length

The experimental results shown in Figure 4, the removal rate (strainaway rate) in length of 2 to 6 increased rapidly, when the length is 6 to 0.8. Then maintained at a smooth state (very slow rise). The reason for this phenomenon is that most of the Chinese words are in the following 3 Chinese characters (6 bytes). Retention rate (rate reserved) was slowly decreased to the lowest point of 0.85.

Based on the experimental results, our system uses 16 bytes as the length of the sliding window. At this time, the removal rate reached 0.87 and retention rate reached 0.975, this is in before the LCS computation significantly removed not similar to document the differences, thereby greatly reducing the computation time, while at the same time for real similar documents TLCS content, only very small.

A. Experiment 3: Evaluating the Effect of the Slope Threshold of the Trusted LCS Region

Random from the public hall download 1000 microblog pages document, they correspond to the similar subset is obtained to assess for each of a selected twitter page document, in its similar subset of algorithms to find the similar twitter page document, and calculate the accurate rate and approximate the recall rate, finally take the average value. By changing the slope threshold from 0.02 to 0.2, the similarity measure is adjusted in the process, so that the approximate recall rate is kept at 0.9, then the corresponding accuracy is calculated. When the slope threshold value is a certain value, the accuracy rate can reach a maximum value.

Experiments show that the accuracy of LCS is improved and the accuracy is improved to some extent by calculating the results of the calculation. If you do not take computing LCS and extract the part of the trusted method is to take the slope threshold for infinite value, accuracy rate will have a certain impact, so take computer LCS and extract the trusted part of the algorithm is of a certain significance.

B. Experiment 4: Comparison with SimHash Algorithm

In the evaluation of the literature, the recall rate is almost the same (recall), simhash Charikar algorithm achieved better accuracy than shingling algorithm (precision). So according to the practice of Henzinger has done some experiments to compare our algorithm and simhash algorithm. We compare the algorithms in two: 1) on the basis of overall accuracy rate and recall rate, accuracy rate of only 2) for the same microblogging and recall.

We realize the simhash algorithm and Henzinger implementation of the difference, this is because we use the standard guidance manual testing of different. We each microblogging document to generate a 128 bit (bit) of the vector, and obtain the threshold number to 97, which in the setting of the recall rate (recall) value and achieved the highest accuracy (precision). In our algorithm, the fingerprint number is 112, the length of the sliding window is 16, the slope of the threshold for the 0.10, resemble rate 0.28 contain rate threshold, the threshold is 0.7. The accuracy rate (precision). The calculation method of the section. Recall (recall) evaluation as follows: We randomly selected from the 430 million document collection of 1000 documents, from where they may be similar with the subset as a test set, a document about 12 thousand. For each document is selected, we calculate the algorithm obtain truly similar documents in the test set (true-similar) and the ratio of simhash algorithm of real similar document. The average value of the ratio is the ratio of our algorithm recall and recall rate of the simhash algorithm.

Experimental results show that our algorithm in accuracy (precision) and recall (recall) showed better: 1) our algorithm in the overall accuracy rate reaches 0.95 and simhash algorithm is 0.72, and our algorithm and simhash algorithm to recall ratio is 1.86; 2) only for the assessment of the same micro Bo documents, our algorithm and simhash find true similar document (true-similar) the number of is almost consistent, and accuracy were 0.91 and 0.52. The reason why

we can get a better accuracy is that we are based on "the content of the document" and not "the matching of the document hash fingerprint", and the influence of the same website template is reduced. Our algorithm can obtain better recall rate because: 1) our algorithm can subsumption relations between two document evaluation and judgment are similar if the condition is met; 2) our algorithm in the face really similar microblogging document but contains different templates more robust, which is why our method can find more on the existence of different microblog page real similar documents.

Observations indicate that false-positive errors of judgment caused by our algorithm mainly comes from three aspects: 1) those describing a similar but different microblogging page document, for example two sales Computer microblogging, describes the same series but different types of computers; 2) microblog template occupy a large proportion of content; 3) microblogging content are small differences, but crucial.

We compared the accuracy and recall rate between the algorithm and the simhash algorithm on a small scale test set. However, because we do not realize the simhash algorithm in large-scale data conditions, so there is no comparison of the efficiency of our algorithm and the simhash algorithm, which will be the future needs of our work to do.

C. Experiment 5: Evaluating the Performance of the Algorithm

Microblog page data is massive data, if the use of this algorithm to calculate the entire data set, the time efficiency is low, it is difficult to achieve the purpose of real-time applications. So we take a crawl Sina microblog a time period of a total of 3000 microblog page similarity detection, time is about 120 seconds, a total of 15% of the similarity of the micro blog page. Experimental results show that the LCS based microblog similar page detection algorithm can not only detect the similar pages of the microblog, but also can be accepted.

VI. CONCLUSION

Through the analysis of the characteristics of micro blog page data, in micro Bo forwarding operation of similar web pages is proposed a page detection method based on LCS microblogging similarity. First is to calculate the similar microblogging page document subset, followed by calculation of the LCS and extract the part of the trusted and final inspection measure microblogging similar pages, ideal effect and high efficiency. Experimental results show that this method can effectively detect similar pages in micro blog data, which is of great significance to reduce the burden of users and improve the efficiency of network public opinion analysis.

However, these still need our further deepening and improvement, for example with the continuous development of micro Bo, the data volume will continue to increase, in the massive data to detect similar pages and efficiency is an eternal proposition; in addition, microblogging is similar to the detection of the page and other methods, it is worthwhile for us to continue to compare and explore and find the most efficient method of the most suitable.

ACKNOWLEDGMENT

This work was supported by a grant from the research projects of Henan province, China (No. 162102210395).

REFERENCES

[1] Shengchen Zhou, Wenting Qu, Yingzi Shi, et al, “Overview on sentiment analysis of chinses microblogging”, Computer applications and software, vol. 30, No. 3, pp.161-163, 2013.

[2] Lin Wang, Shi Feng, Weili Xu, “A filtering approach for spam discrimination and content similarity double detection for microblog text stream”, Computer applications and software, vol. 29, No. 8, pp. 25-29, 2012.

[3] S. Liuzzo, P. Tomei. “A global adaptive learning control for robotic manipulators”, Automatica, vol. 44, No. 5, pp.1379–1384, 2008.

[4] Yinglong Wang, Bingru Yang, Zefeng Song, et al. “LCS algorithm research for gene sequence similar degree”, Computer engineering and applications, vol. 43, No.31, pp. 45-47, 2007.

[5] Yang Zhao,Yanguang Cai. “A novel local exploitation scheme for conditionally breeding real-coded genetic algorithm”, Multimedia Tools and Applications, DOI 10.1007/s11042-016-3493-0. 2016.

[6] E. Myers. “An O(ND) difference algorithm and its variations”, Algorithmica, vol. 1, No. 2, pp. 251–266, 1986.

[7] Yang Zhao, Defu Cheng, XiaoJun Yang. “Approximation solutions for local fractional Schrödinger equation in the one-dimensional cantorian system”, Advances in Mathematical Physics, vol. 13, No. 03, pp.1-5, 2013.

[8] Yang Zhao, Dumitru Baleanu, Carlo Cattani et al. “Maxwell’s equations on Cantor sets: a local fractional approach”, Advances in High Energy Physics, vol. 12, No. 02, pp.1-9, 2013.

[9] A. Z. Broder, S. C. Glassman, M. S. Manasse et al. “Syntactic clustering of the web. computer networks”, Computer networks & isdn systems, vol. 29, No. 8-13, pp.1157-1166, 1997.